Combining Word Embedding and Lexical Database for Semantic Relatedness Measurement

Yang-Yin Lee, Hao Ke, Hen-Hsen Huang, Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{yylee, hke, hhhuang}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

ABSTRACT
While many traditional studies on semantic relatedness utilize the lexical databases, such as WordNet[1] or Wikitionary[2], the recent word embedding learning approaches demonstrate their abilities to capture syntactic and semantic information, and outperform the lexicon-based methods. However, word senses are not disambiguated in the training phase of both Word2Vec and GloVe, two famous word embedding algorithms, and the path length between any two senses of words in lexical databases cannot reflect their true semantic relatedness. In this paper, a novel approach that linearly combines Word2Vec and GloVe with the lexical database WordNet is proposed for measuring semantic relatedness. The experiments show that the simple method outperforms the state-of-the-art model SensEmbed.

Keywords: Semantic relatedness; word embedding; WordNet; GloVe; Word2Vec.

1. INTRODUCTION
Semantic relatedness, which computes the association degree of two objects such as words, entities and texts, is fundamental for many applications. It has long been thought that when human measure the relatedness between a pair of words, a deeper reasoning is triggered to compare the concepts behind the words. While many semantic relatedness researches in the past utilized lexical databases such as WordNet and Wikitionary, the recent word embedding approaches have demonstrated their abilities to capture both syntactic and semantic information [4, 5]. Among the embedding representations, Word2Vec and GloVe are widely adopted for many researches. However, word senses are not disambiguated in the training phase of both Word2Vec and GloVe. That affects the measurement of semantic relatedness. On the other way round, WordNet and Wikitionary are well-structured ontology that provides senses of each word. However, the path length between any two senses of words cannot reflect their true semantic relatedness (e.g., some word pairs have the same path distance, but different human rated relatedness scores).

In this paper we propose a novel approach for measuring semantic relatedness by joining word embedding and lexical database via linear combination. The experimental results show that the simple and efficient method outperforms the state-of-the-art model SensEmbed [3] when we apply a standardization process to GloVe.

[1] https://wordnet.princeton.edu/
[2] https://www.wiktionary.org/

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2889395.

2. OUR APPROACH
The first part of our approach is to compute the cosine similarity between a pair of words represented by dense vector embeddings of words. However, the representation is unique for each word even though it is ambiguous, i.e., it has more than one sense. For example, the word *bass* has at least two senses: (1) {bass, bass part}, the lowest part in polyphonic music; and (2) {bass}, the lowest part of the musical range. The words in the brackets have the same sense under some circumstance. To overcome the drawbacks of the aforementioned models, we adopt a lexical database to aid the semantic relatedness measurement.

In the experiments, WordNet is selected as our lexical database. Of course, other lexical databases are also applicable. Let $S_{i,m}$ be the m-th sense associated to the word w_i. The path distance between the senses of any two words can be computed in the WordNet. For example, the path distance of the synset *cat.n.01* (the first sense of noun cat) and *dog.n.01* (the first sense of noun dog) is 5. If the path distance between two synsets is smaller, then they should be more similar to each other. The final semantic relatedness score of two words which combines word embedding and lexical database is defined as follows.

$$\mathrm{rel}(w_i, w_j) = \max_{m,n} \lambda \cos\left(v_{w_i}, v_{w_j}\right) + (1-\lambda)\frac{1}{dist(S_{i,m}, S_{j,n})} \quad (1)$$

where $\mathrm{dist}(S_{i,m}, S_{j,n})$ is the distance between the two senses $S_{i,m}$ and $S_{j,n}$. v_{w_i} and v_{w_j} are the vector representations of word w_i and w_j in the word embedding. λ is a weighting factor.

3. EXPERIMENTS
3.1 Datasets and Experimental Setup
We downloaded four benchmark datasets from the web: RG-65 [6], WordSim353 [2], YP130 [7], and MEN [1]. The WS353-sim and the WS353-rel are the subsets of the WordSim353, for measuring the word similarity and the word relatedness tasks, respectively. We adopted two word embeddings in our experiments: GloVe and Word2Vec. GloVe is trained on the Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors).

Word2Vec model is trained on part of Google News dataset with about 100 billion words. Besides directly applying the vector representation of the word embedding models, we also performed the standardization process (z-[word embedding]). The standard score transforms each dimension of the word vectors to have zero mean and unit variance using the following formula $\mathbf{z} = \frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are derived from the word embedding model. Since we found that many words in both word embeddings do not appear in the WordNet. We also considered another

Table 1. Spearman (ρ) and Pearson (r) correlation of different semantic relatedness measures on RG-65, WS353-all, WS353-sim, WS353-rel, YP130 and MEN datasets.

Method	RG-65	WS353-all	WS353-sim	WS353-rel	YP130	MEN
	ρ/r	ρ/r	ρ/r	ρ/r	ρ/r	ρ/r
SensEmbed$_{closet}$	0.894/None	0.714/None	0.756/None	0.645/None	0.734/None	0.779/None
SensEmbed$_{weighted}$	0.871/None	0.779/None	0.812/None	0.703/None	0.639/None	0.805/None
Word2Vec	0.761/0.772	0.694/0.649	0.777/0.768	0.622/0.583	0.570/0.589	0.782/0.770
z-Word2Vec	0.758/0.773	0.693/0.650	0.777/0.769	0.621/0.582	0.570/0.588	0.782/0.769
Word2Vec + path	0.873/0.830	0.707/0.656	0.812/0.793	0.622/0.584	0.731/0.782	0.793/0.770
z-Word2Vec + path	0.874/0.830	0.707/0.656	0.810/0.797	0.622/0.583	0.736/0.783	0.794/0.770
z$_{sub}$-Word2Vec + path	0.883/0.834	0.699/0.658	0.810/0.790	0.612/0.593	0.747/0.787	0.803/0.781
GloVe	0.817/0.800	0.632/0.639	0.698/0.704	0.571/0.603	0.502/0.467	0.744/0.742
z-GloVe	0.823/0.815	0.678/0.679	0.736/0.742	0.632/0.653	0.534/0.525	0.772/0.768
GloVe + path	0.903/0.867	0.653/0.653	0.786/0.754	0.571/0.603	0.719/0.760	0.750/0.746
z-GloVe + path	0.910/0.868	0.721/0.697	0.821/0.786	0.663/0.667	0.732/0.777	0.792/0.783
z$_{sub}$-GloVe + path	**0.916**/0.866	**0.788**/**0.734**	**0.848**/**0.815**	**0.747**/**0.733**	**0.749**/0.788	**0.828**/**0.813**
z$_{sub}$-GloVe+path (λ=0.75)	0.908/**0.877**	0.777/0.704	**0.848**/**0.815**	0.710/0.687	0.745/**0.789**	0.817/0.781

standardization method (z$_{sub}$-[word embedding]): $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are calculated using the vocabulary from the intersection of the word embedding and the WordNet. For each dataset, the weighted parameter λ is set from 0 to 1 with 0.05 as the step size. We also compared our model to one of the state-of-the-art approach SensEmbed [3], which combines a sense embedding and a knowledge base. Figure 1 shows Spearman correlation on the RG-65 dataset with z$_{sub}$-GloVe+path under different λs. SensEmbed is compared. When λ is in the range of 0.45-0.85, our model outperforms SensEmbed. The λ settings are similar in the other datasets.

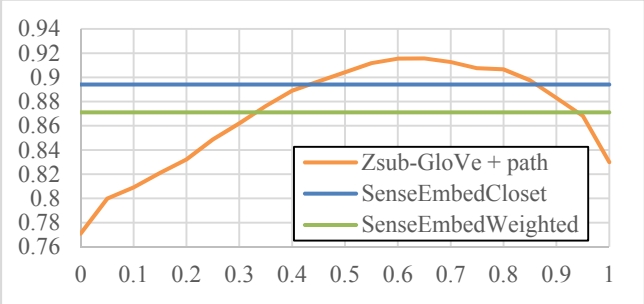

Figure 1. Selection of λ

3.2 Results and Discussion

Table 1 shows the complete experimental results of our method together with the SensEmbed approach. We report the results in Spearman (ρ) and Pearson (r) correlation coefficient. From Table 1 we can find that almost all tasks benefit from the combination of the word embedding and the WordNet (e.g., Word2Vec vs. Word2Vec+path, and z-GloVe vs. z-GloVe+path). It is clear that z$_{sub}$-GloVe+path has the best performance among all models and outperforms the SensEmbed in every task's ρ. A fixed λ version (λ=0.75) listed in the last row shows the robustness of our approach. In Word2Vec, the standardization process does not affect the model's performance. Comparatively, the standardization process in GloVe improves the performance in every task. The reason may be that Word2Vec's training proceeds in a stochastic fashion that attempts to maximize the log probability, while GloVe factorizes the word-word co-occurrence matrix. Finally, the z$_{sub}$ process further improves the performance (compares to z-) in the GloVe, but not in the Word2Vec. This finding argues that the words in GloVe, but not appearing in the WordNet might not be trained well.

4. CONCLUSIONS

We propose a novel yet simple approach to measure the semantic relatedness with linearly combining the word embedding models Word2Vec and GloVe, and a lexical database WordNet. One benefit of our model is that our approach is highly adaptable since our model is capable of adopting different word embeddings or different lexical databases. We evaluate our method on six tasks, including RG-65, WS353-all, WS353-sim, WS353-rel, YP130 and MEN. The experimental results show that the path information is indeed beneficial to the semantic relatedness measurement. Another point worthy of mentioning is that the GloVe model is sensitive to the standardization. Different standardization process may have a large impact on the final model.

5. ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-102-2221-E-002-103-MY3 and MOST-104-2221-E-002-061-MY3.

6. REFERENCES

[1] Bruni, E. et al. 2014. Multimodal Distributional Semantics. *J. Artif. Intell. Res.(JAIR)*. 49, (2014), 1–47.

[2] Finkelstein, L. et al. 2001. Placing search in context: The concept revisited. *Proceedings of World Wide Web*, 406–414.

[3] Iacobacci, I. et al. 2015. SensEmbed: learning sense embeddings for word and relational similarity. *Proceedings of ACL* (2015), 95–105.

[4] Mikolov, T. et al. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (2013), 3111–3119.

[5] Pennington, J. et al. 2014. Glove: Global vectors for word representation. *Proceedings of EMNLP*, 1532–1543.

[6] Rubenstein, H. and Goodenough, J.B. 1965. Contextual correlates of synonymy. *CACM*. 8, 10 (1965), 627–633.

[7] Yang, D. and Powers, D.M. 2005. Measuring semantic similarity in the taxonomy of WordNet. *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38* (2005), 315–322.